Mining Complex Data Generated by Collaborative Platforms

Dmitry I. Ignatov[1], Alexandra Yu. Kaminskaya[1,2], Anastasya A. Bezzubtseva[1,2], Ekaterina L. Chernyak[1,2], Konstantin N. Blinkin[1], Daniil R. Nedumov[1], Olga N. Chugunova[1], Andrey V. Konstantinov[1], Nikita S. Romashkin[1] Fedor V. Strok[1], Daria A. Goncharova[1,2], Jonas Poelmans[1], and Rostislav E.Yavorsky[2]

[1] National Research University Higher School of Economics, Russia, 101000, Moscow, Myasnitskaya str., 20
dignatov@hse.ru
http://www.hse.ru
[2] Witology
rostislav.yavorskiy@witology.com
http://www.witology.com

Abstract. In a crowdsourcing project several participants discuss and solve one common problem, propose their ideas, evaluate ideas of each other, etc. We propose the novel instrument CrowDM for analyzing data generated by collaborative platforms. The initial version of the system combines several innovative techniques for structured and unstructured data analysis. Formal Concept Analysis, multimodal clustering and association rule mining are the key instruments for identifying patterns in object-oriented data. Keyword and colocation extraction methods are also included for mining unstructured texts. We first describe the overall methodology underlying CrowDM and then showcase results of initial experiments on data obtained from the company Witology.

Keywords: Collaborative and Crowdsourcing Platforms, Data Mining, Formal Concept Analysis, Multimodal Clustering.

1 Introduction

Several years ago the Russian crowsourcing companies Witology [1] and Wikivote [2] were founded. They were following in the footsteps of their succesful predecessors in the USA (e.g. Spigit [3], BrightIdea [4] and InnoCentive [5]) and Europe (Imaginatik [6]). All of these companies heavily rely on collaborative platforms for completing their projects. Recently several Russian projects using collaborative technologies were completed succesfully, including Sberbank-21 and its National Entrepreneurial Initiative-2012 [7].

At the basis of a collaborative platform is a socio-semantic network [8,9,10,11] which generates a lot of data for analysis. While participating in a project, users of such a crowdsourcing platform [12] can discuss and solve one common problem, propose their ideas, evaluate ideas of each other, etc. From the discussions

between users and the ranking of their ideas we can easily obtain the most popular ideas and the users who generated them. If we however want to gain a deeper understanding of the behavior of users, develop more adequate and objective ranking criteria, perform dynamic and complex statistical analyses, etc. more sophisticated methods are needed. In order to go beyond discovering the fool's gold, traditional text mining, cluster and community detection methods need to be adapted or even fully redesigned.

In this paper we propose the collaborative platform data analysis system Crowd Data Mining (CrowDM). We discuss its architecture as well as the data analysis methods it offers. We describe in detail how keywords can be extracted from data resulting from a crowdsourcing project and analyzed with Formal Concept Analysis (FCA) [13,14], biclustering, etc.

The remainder of the paper is organized as follows. In section 2 we describe the essentials of FCA theory, biclustering, keyword extraction and peculiarities of the Witology data. In section 3 we discuss the analysis scheme of the developed system. In section 4 we present the results of our first experiments with the Sberbank-21 data. Section 5 concludes our paper and describes some possible directions for future research.

2 Mathematical models and methods

At the initial stage of collaborative platform data analysis two data types were identified: data without using keywords (links, evaluations, user actions) and data with keywords (all user-generated content). These two data types totally correspond with two components of a socio-semantic network. For the analysis of the 1st type of data (with keywords) we suggest to apply Social Network Analysis (SNA) methods, clustering (biclustering and triclustering [15,16,17], spectral clustering), FCA (concept lattices, implications, association rules) and its extensions for multimodal data, triadic, for instance [18]; recommender systems [19,20,21] and statistical methods of data analysis [22] (the analysis of distributions and average values). Methods described in this paper are colored blue at the analysis scheme (see fig. 2).

2.1 Formal Concept Analysis

The protagonists of crowdsourcing projects (and corresponding collaborative platforms) are platform users (project participants). We consider them as *objects* for analysis. More than that, each object can (or cannot) possess a certain set of *attributes*. The user's attributes can be: topics which the user discussed, ideas which he generated or voted for, or even other users. The main instrument for analysis of such object-attribute data is FCA. Let us give formal definitions. *The formal context* in FCA is a triple $\mathbb{K} = (G, M, I)$, where G is a *set of objects*, M is a *set of attributes*, and the relation $I \subseteq G \times M$ shows which object possesses which attribute. For any $A \subseteq G$ and $B \subseteq M$ one can define *Galois operators*:

$$A' = \{m \in M \mid gIm \text{ for all } g \in A\}, \quad (1)$$
$$B' = \{g \in G \mid gIm \text{ for all } m \in B\}.$$

The operator $''$ (applying the operator $'$ twice) is a *closure operator*: it is idempotent ($A'''' = A''$), monotonous ($A \subseteq B$ implies $A'' \subseteq B''$) and extensive ($A \subseteq A''$). The set of objects $A \subseteq G$ such that $A'' = A$ is called closed. The same is for closed attribute sets, i.e. subsets of a set M. A couple (A, B) such that $A \subset G$, $B \subset M$, $A' = B$ and $B' = A$, is called *formal concept* of a context $\mathbb{K}$. The sets A and B are closed and called *extent* and *intent* of a formal concept (A, B) correspondingly. For the set of objects A the set of their common attributes A' describes the similarity of objects of the set A, and the closed set A'' is a cluster of similar objects (with the set of common attributes A'). The relation "to be more general concept" is defined as follows: $(A, B) \geq (C, D)$ iff $A \subseteq C$. The concepts of a formal context $\mathbb{K} = (G, M, I)$ ordered by extensions inclusion form a lattice, which is called *concept lattice*. For its visualization the *line diagrams* (Hasse diagrams) can be used, i.e. cover graph of the relation "to be more general concept". In the worst case (Boolean lattice) the number of concepts is equal to $2^{\{\min |G|, |M|\}}$, thus, for large contexts, FCA can be used only if the data is sparse. Moreover, one can use different ways of reducing the number of formal concepts (choosing concepts by stability index or extent size).

2.2 Biclustering

An alternative approach is a relaxation of the definition of formal concept as maximal rectangle in an object-attribute matrix which elements belong to the incidence relation. One of such relaxations is a notion of object-attribute bicluster [16]. If $(g, m) \in I$, then (m', g') is called object-attribute bicluster with the density $\rho(m', g') = |I \cap (m' \times g')|/(|m'| \cdot |g'|)$.

The main features of OA-biclusters are listed below:

1. For any bicluster $(A, B) \subseteq 2^G \times 2^M$ it is true that $0 \leq \rho(A, B) \leq 1$.
2. OA-bicluster (m', g') is a formal concept iff $\rho = 1$.
3. If (m', g') is a bicluster, then $(g'', g') \leq (m', m'')$.

Let $(A, B) \subseteq 2^G \times 2^M$ be a bicluster and ρ_{min} be a non-negative real number such that $0 \leq \rho_{\min} \leq 1$, then (A, B) is called *dense*, if it fits the constraint $\rho(A, B) \geq \rho_{\min}$. The above mentioned properties show that OA-biclusters differ from formal concepts since unit density is not required. Graphically it means that not all the cells of a bicluster must be filled by a cross (see fig. 1). Besides formal lattice construction and visualization by means of Hasse diagrams one can use implications and association rules for detecting attribute dependencies in data. Then, using the obtained results, it is easy to form recommendations (for example, offering users the most interesting discussions for them). Furthermore, structural analysis can be performed and then used for finding communities.

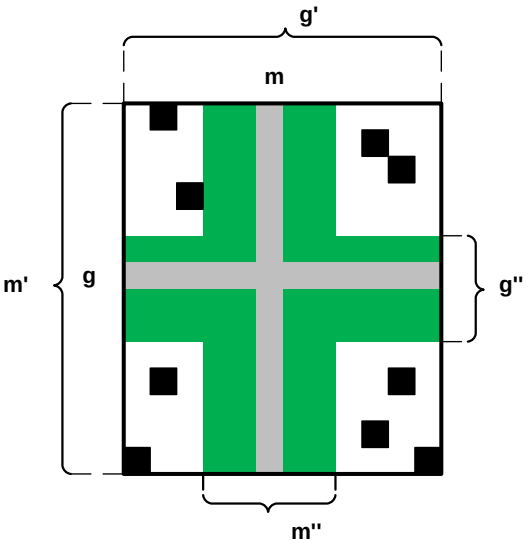

Fig. 1. OA-bicluster.

Statistical methods are helpful for frequency analysis of the different users' activities. Almost all of the above mentioned methods can be applied to data containing users' keywords (in this case they become attributes of a user).

2.3 Keywords and keyphrases extraction

We consider *Keywords (keyphrases)* as a set of the most significant words (phrases) in a text document that can provide a compact description for the content and style of this document. In the remainder of this paper we do not always differentiate between keywords and keyphrases, assuming that a keyword is a particular case of a keyphrase. In our project two similar problems of keywords and keyphrases extraction arise:

1. keywords and keyphrases of the whole Witology forum;
2. keywords and keyphrases of one user, topic etc.

In the first case we concentrate on finding syntactically well associated keywords (keyphrases). In the second case specific words and phrases of a certain user or topic are the subject of interest. Hence, we have to use two different methods for each keyword (keyphrase) extraction problem. The first one is solved by using any statistical measure of association, such as Pointwise Mutual Information (PMI), T-Score or Chi-Square [23]. To solve the second problem we may use TF-IDF or Mutual Information (MI) measures that reflect how important the word or phrase is for the given subset of texts. All the above mentioned measures define the weight of a specific word or phrase in the text. The words and phrases

of the highest weight then can be considered as keywords and keyphrases. We are more interested in the quality of extracted keywords and keyphrases than in the way we obtain them. To tokenize texts we use a basic principle of word separation: there should be either a space or a punctuation mark between two words. A hyphen between two sequences of symbols makes them one word. To lemmatize words we use Russian AOT lemmatizer [24], which is far from being ideal, but it is the only freely available one (even for commercial usage) for processing Russian texts. To normalize bi- and tri-grams we use one of our Python scripts that normalizes phrases according to their formal grammatical patterns. We are going to use formal contexts based on sets of extracted keyphrases and people who use them, the occurrence of keyphrases in texts and so on. By analogy, keyphrases, texts and users all together form a tricontext for further analysis. Moreover, keyphrases are an essential part of a socio-semantic network model, where they are used for semantic representation of the network's nodes.

3 Analysis scheme

The data analysis scheme of CrowDM, which is developed now by the project and educational team of Witology and National Research University Higher School of Economics is presented in figure 2. As it was mentioned before, after downloading data from a platform database, we obtain formal contexts and text collections. In turn, the latter become formal contexts as well after keyword extraction. After that, the resulting contexts are analyzed.

4 First experiments results

For carrying out experiments we constructed formal concepts where objects are users of the platform and attributes are ideas which users proposed within one of 5 project topics ("Сбербанк и частный клиент" ("Sberbank and private client")). We selected only the ideas that reached the end or almost the end of the project. An object"user" has an attribute "idea" if this user somehow contributed to the discussion of this idea, i.e. he is an author of the idea, commented on the idea and evaluated the idea or comments which were added to the idea. Thus, the extracted formal concepts (U, I), where U is a set of users, I is a set of ideas, correspond to so called epistemic communities (communities of interests), i.e. the set of users U who are interested in the ideas of I. Figure 3 displays the diagram of the obtained concept lattice.

Each node of the diagram coincides with one formal concept (in total the lattice contains 198 concepts). A node is marked by the label of an object or an attribute if this object (moving bottom-up by diagram) or attribute (moving top-down) first appeared in this node. It is obvious that the obtained diagram is too awkward to be analyzed as a static image. Usually in such cases one can use order filters or diagrams of the sets of stable concepts or iceberg-lattices for visualization. We will showcase how to read a concept lattice using the lattice fragment in figure 4. The experiments were carried out using the program

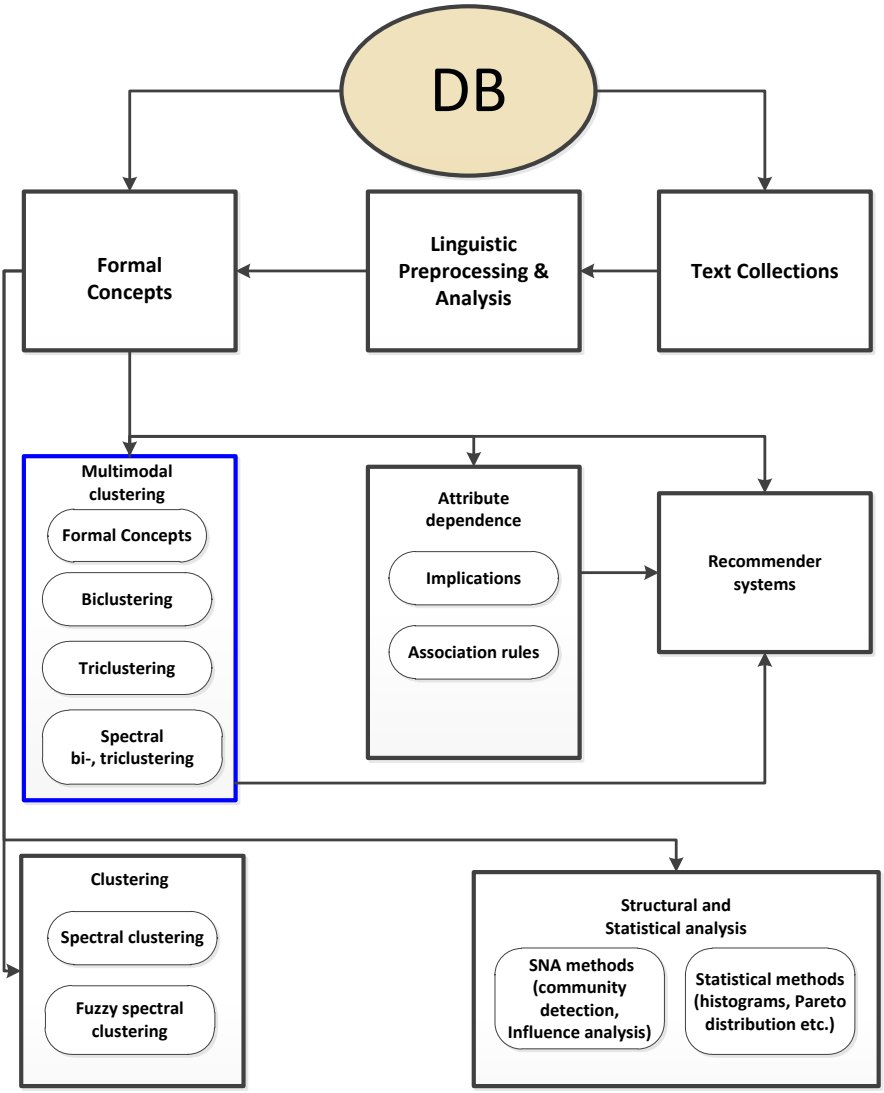

Fig. 2. The data analysis scheme of CrowDM (all indicated methods is implemented by the paper's authors.

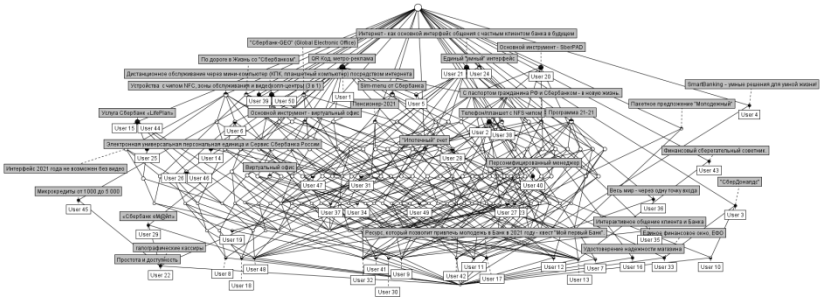

Fig. 3. Concept lattice diagram for users-ideas context (labels are given in Russian).

Concept Explorer (ConExp) which was developed for applying FCA algorithms to object-attribute data [25]. Clicking on a lattice node, one can see the objects and attributes corresponding to the concept which this node represents. Objects are accumulated from below (in the given example the set of objects contains User45 and User22), attributes come from above (we have only one attribute, "Микрокредиты от 1000 до 5000"("Microcredits from 1000 to 5000")). This means that User45 and User22 together took part in the discussion of the given idea and nobody else discussed it.

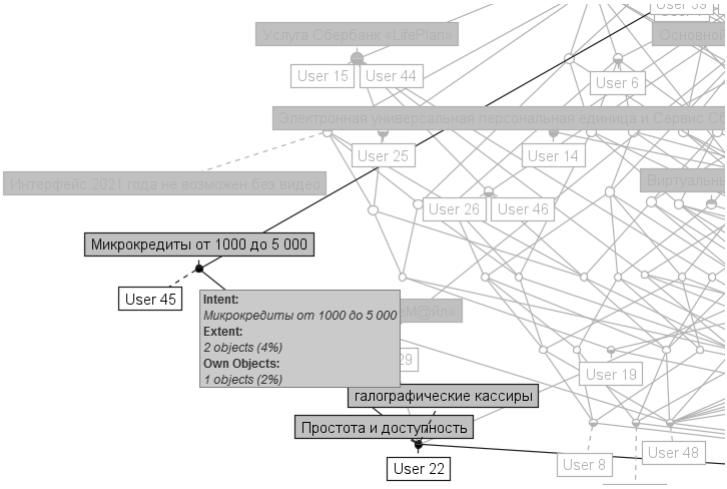

Fig. 4. Fragment of concept lattice diagram (labels are given in Russian)

We demonstrate the results of applying biclustering algorithms on the same data below.

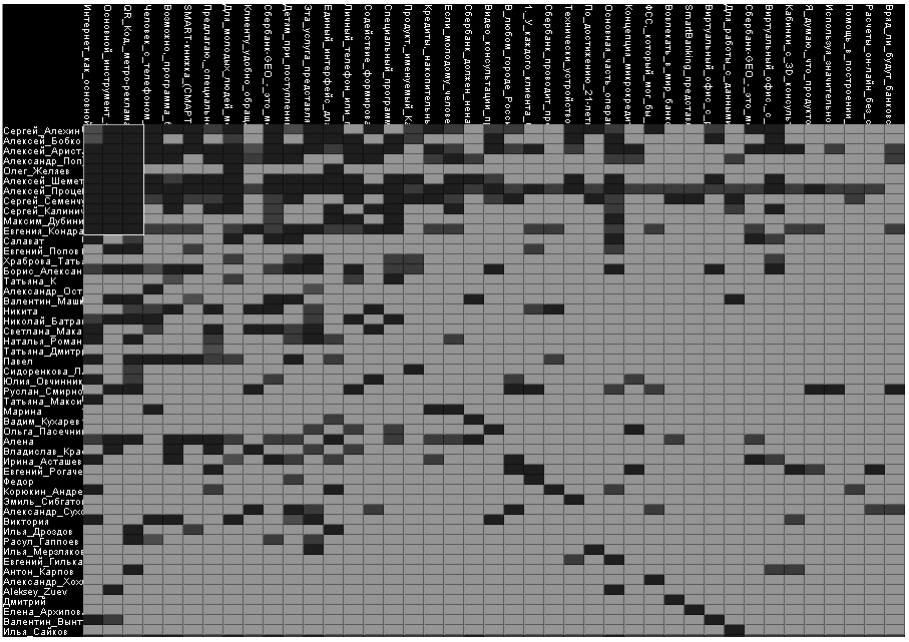

Fig. 5. Results of biclustering algorithm BiMax

Let us explain the figure 5. During experiments we used the system for gene expression data analysis BicAT [15]. Rows correspond to users, columns are ideas of a given topic (e.g., "Sberbank and private client" or "Сбербанк и частный клиент" in Russian), in the discussion of which users participated. The color of the cell of the corresponding row and column intersection depicts the contribution intensity of a given user to a given idea. The contribution is a weighted sum of the number of comments and evaluations to that idea and takes into account the fact whether this user is an author of this idea. The lightest cells coincide with zero contribution, the brightest ones (fig. 6, top left cell) show the maximum contribution; here the columns are ideas and rows are users. After data discretization (0 – zero contribution, 1 – otherwise) we applied the BiMax algorithm which found some biclusters (see fig. 6 for example). Since one of the important crowdsourcing project problems is the search of people with similar ideas, the presented bicluster with 11 users is most interesting while other found biclusters contained 4-5 users on average (we constrained the number of ideas in a bicluster to be strictly greater than 2).

Then, to gain a better understanding of the evaluation process in the project, evaluation distribution was plotted in several ways. One of them is presented in fig. 7; it shows the cumulative number of users, who made more than a certain amount of evaluations during the entire project.

The horizontal axis displays the amount of submitted evaluations. The vertical axis represents the number of users, who made more than a fixed amount

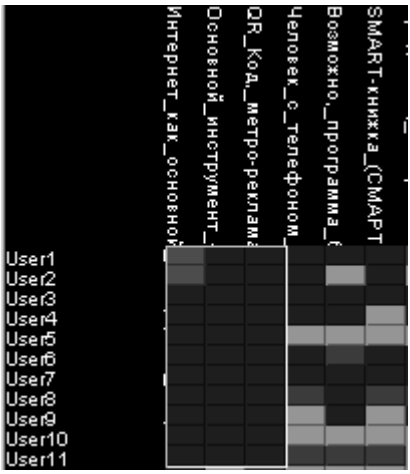

Fig. 6. Biclsuter with a large number of users

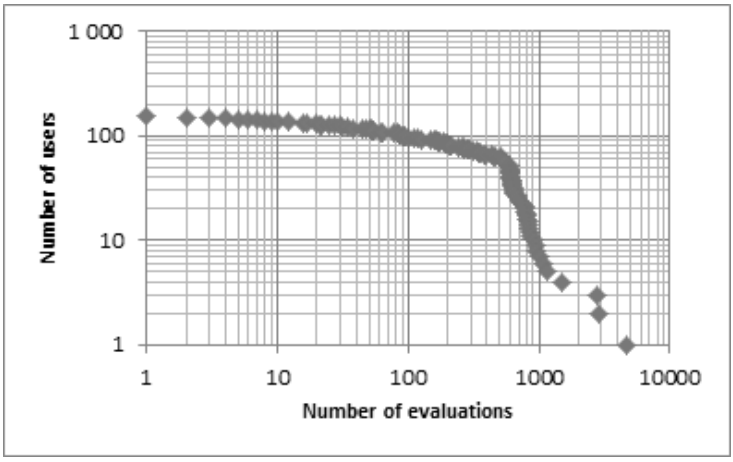

Fig. 7. Evaluation distribution

of evaluations. For instance, there is only one participant, who produced more than 5000 evaluations, and one more person, who made more than 3000 but less than 5000 evaluations. Thus, the rightmost dot on the X-axis shows the first participant (the y-coordinate is 1), and the next dot shows both of them (the y-coordinate is 2). The total number of users, who have once evaluated something, is 167. The set of graph points is explicitly split into two parts: the long gentle line (from $x = 0$ to 544 inclusive) and the steep tail. The fact, that both lines seem almost straight in logarithmic scales, indicates that the evaluation activity on the project might follow a Pareto distribution. It is reasonable to seek the individual distribution functions for the main and the tail parts of the sample, as testing the whole sample for goodness of fit to a Pareto distribution results in strong rejection of the null hypothesis ($H0$: "The sample follows a Pareto distribution"). This analysis implies useful consequences according to the well-known "80:20" rule:

$$W = P^{(\alpha-2)/(\alpha-1)},$$

which means that the fraction W of the wealth is in the hands of the richest P of the population. In our case for $\alpha = 3.41$ (the steep tail), i.e. 69% of users make 80% of all idea evaluations, there is no traditional disproportion, but for the first part (from $x = 0$ to 544 inclusive) this formula is inapplicable ($\alpha = 1.48$).

5 Conclusion

The results of our first experiments suggest that the developed methodology will be useful for analysis of collaborative systems data and resource-sharing systems. The most important directions for future work include the analysis of textual information generated by users, applying multimodal clustering methods and using them for developing recommender systems.

Acknowledgments. The work was performed by the scientific-educational group "Algorithms of Data Mining for Innovative Projects Internet Forum" of National Research University Higher School of Economics.

References

1. Witology company, http://witology.com/
2. Wikivote company, http://www.wikivote.ru/
3. Spigit company, http://spigit.com/
4. Brightidea company, www.brightidea.com/
5. Innocentive comp, http://www.innocentive.com/
6. Imaginatik company, http://www.imaginatik.com/
7. Sberbank-21, national entrepreneurial initiative-2012, http://sberbank21.ru/
8. Roth, C.: Generalized preferential attachment: Towards realistic socio-semantic network models. In: ISWC 4th Intl Semantic Web Conference, Workshop on Semantic Network Analysis, Galway, Ireland,. Volume 171 of CEUR-WS Series (ISSN 1613-0073). (2005) 29–42

9. Cointet, J.P., Roth, C.: Socio-semantic dynamics in a blog network. In: CSE (4), IEEE Computer Society (2009) 114–121
10. Roth, C., Cointet, J.P.: Social and semantic coevolution in knowledge networks. Social Networks **32** (2010) 16–29
11. Yavorsky, R.: Research Challenges of Dynamic Socio-Semantic Networks. In Ignatov, D., Poelmans, J., Kuznetsov, S., eds.: CEUR Workshop proceedings Vol-757, CDUD'11 - Concept Discovery in Unstructured Data. (2011) 119–122
12. Howe, J.: The rise of crowdsourcing. Wired (2006)
13. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. 1st edn. Springer-Verlag New York, Inc., Secaucus, NJ, USA (1999)
14. Poelmans, J., Ignatov, D., Viaene, S., Dedene, G., Kuznetsov, S.: Text mining scientific papers: a survey on fca-based information retrieval research. In Perner, P., ed.: 12th Industrial Conference on Data Mining, July 13-20, Berlin, Germany. Volume 7377 of Lecture Notes in Artificial Intelligence., Springer (2012) 273–287
15. Barkow, S., Bleuler, S., Prelic, A., Zimmermann, P., Zitzler, E.: Bicat: a biclustering analysis toolbox. Bioinformatics **22**(10) (2006) 1282–1283
16. Igantov, D.I., Kaminskaya, A.Y., Kuznetsov, S., Magizov, R.A.: Method of Biclusterzation Based on Object and Attribute Closures. In: Proc. of 8-th international Conference on Intellectualization of Information Processing (IIP 2011). Cyprus, Paphos, October 17–24, MAKS Press (2010) 140–143 (in Russian).
17. Ignatov, D.I., Kuznetsov, S.O., Magizov, R.A., Zhukov, L.E.: From Triconcepts to Triclusters. In: Proceedings of the 13th international conference on Rough sets, fuzzy sets, data mining and granular computing. RSFDGrC'11, Berlin, Heidelberg, Springer-Verlag (2011) 257–264
18. Jäschke, R., Hotho, A., Schmitz, C., Ganter, B., Stumme, G.: TRIAS–An Algorithm for Mining Iceberg Tri-Lattices. In: Proceedings of the Sixth International Conference on Data Mining. ICDM '06, Washington, DC, USA, IEEE Computer Society (2006) 907–911
19. Ignatov, D.I., Kuznetsov, S.O.: Concept-based Recommendations for Internet Advertisement. In Belohlavek, R., Kuznetsov, S.O., eds.: Proc. CLA 2008. Volume Vol. 433 of CEUR WS., PalackΓS University, Olomouc, 2008 (2008) 157–166
20. Ignatov, D., Poelmans, J., Zaharchuk, V.: Recommender System Based on Algorithm of Bicluster Analysis RecBi. In Ignatov, D., Poelmans, J., Kuznetsov, S., eds.: CEUR Workshop proceedings Vol-757, CDUD'11 - Concept Discovery in Unstructured Data. (2011) pp. 122–126
21. Ignatov, D.I., Poelmans, J., Dedene, G., Viaene, S.: A New Cross-Validation Technique to Evaluate Quality of Recommender Systems. In Kundu, M.K., Mitra, S., Mazumdar, D., Pal, S.K., eds.: PerMIn. Volume 7143 of LNCS., Springer (2012) 195–202
22. Clauset, A., Shalizi, C.R., Newman, M.E.J.: Power-law distributions in empirical data. SIAM Rev. **51**(4) (November 2009) 661–703
23. Manning, C.D., Schütze, H.: Foundations of statistical natural language processing. MIT Press, Cambridge, MA, USA (1999)
24. Russian project on automatic text processing, www.aot.ru
25. Grigoriev, P., Yevtushenko, S.: Elements of an agile discovery environment. In Grieser, G., Tanaka, Y., Yamamoto, A., eds.: Discovery Science. Volume 2843 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg (2003) 311–319